# ILLUSTRATIVE EXAMPLES OF CLUSTERING USING THE MIXTURE METHOD AND TWO COMPARABLE METHODS FROM SAS

K.E. BASFORD, W.T. FEDERER AND N.J. MILES-MCDERMOTT

BU-921-M January 1987

* Partially supported by Mathematical Sciences Institute and by the Australian-American Education Foundation.

ABSTRACT

The technique of clustering uses the measurements on a set of elements to identify clusters or groups of elements such that there is relative homogeneity within the groups and heterogeneity between the groups. In the associated technical report '86-38 and BU-920-M, "Introduction to the use of mixture models in clustering" by K.E. Basford, the mixture model approach is explained in detail and discussed in relation to other clustering techniques. Under this approach to clustering, the elements are assumed to be a sample from a mixture of several populations in various proportions. The practical application to two real data sets is considered here with the density function in each underlying population assumed to be normal. To provide a base for comparison, two SAS clustering methods with similar assumptions are also considered. The data are analysed using:

KMM - Normal mixture model method,

SAS (CLUSTER) - Ward's method, and

SAS (CLUSTER) - EML method;

the results are discussed.

## 1. INTRODUCTION

This work follows on from technical report '86-38 and BU-920-M, "Introduction to the use of mixture models in clustering" by K.E. Basford. There the reasons for the recent emphasis on this model based approach are discussed. Also, the formal definition of the mixture maximum likelihood method of clustering is given.

In this report, the practical application to two real data sets is considered. Initially, the mixture model approach is again defined, concentrating on the case where the underlying parametric form is the normal distribution. Some of the difficulties associated with its application are discussed. Then to provide a base for comparison, two similar methods of clustering in SAS, Ward's method and the EML method, were chosen because of their similarity in assumptions to the mixture maximum likelihood method. The assumptions for the methods are stated and the differences between these hierarchical techniques and the mixture model approach are clearly stated. In the next section, the data sets chosen for illustration are explained. Finally, the results of applying these methods of cluster analysis are reported and discussed for each data set.

## 2. MIXTURE MAXIMUM LIKELIHOOD APPROACH

The technique of clustering uses the measurements on a set of elements to identify clusters or groups of elements, such that there is relative homogeneity within the groups and heterogeneity between the groups. Under the mixture maximum likelihood approach, it is assumed that the elements are sampled from a mixture of several populations in various proportions. Estimates of the distributions of the underlying populations can then be obtained using the likelihood principle, and the elements can be allocated to these populations on the basis of their estimated posterior probabilities. The mixture method is model based, in that the form of the density of an observation in each of the underlying populations has to be specified. A common approach, and the only one considered in this report, is to take the component densities to be multivariate normal. Even if the estimates of the parameters are not reliable, some empirical studies (Hernandez-Avila, 1979) suggest that the mixture method applied with normal component densities may be fairly robust from the clustering view-point of being able to separate data in the presence of multimodality. This is further supported by Basford (1985) and Basford and McLachlan (1985a, b, c, and d).

Using the notation introduced in the associated technical report '86-38 and, BU-920-M, let $\underset{\sim}{x}_1,\cdots,\underset{\sim}{x}_n$ denote the observed values of a random sample of n p-dimensional observations taken from a mixture of a specified number, say

g, of underlying populations $\Pi_1,\cdots,\Pi_g$. The proportions in which the populations are represented in the mixture are unknown, and will be denoted by $\underset{\sim}{\pi} = (\pi_1,\cdots,\pi_g)'$. Let the density of an observation $\underset{\sim}{x}$ from $\Pi_i$ be given by $f_i(\underset{\sim}{x};\ \underset{\sim}{\nu})$ where $\underset{\sim}{\nu}$ denotes the vector of unknown population parameters. The mixture method of clustering can be applied, at least in principle, provided the form of these densities is known; see the discussion and references in Basford (1986). However, the most widely studied examples of this formulation concern random samples from a mixture of normal distributions.

An observation $\underset{\sim}{x}$ in the superpopulation $\Pi$ has the mixture density given by

$$f(\underset{\sim}{x};\ \underset{\sim}{\nu},\underset{\sim}{\pi}) = \sum_{i=1}^{g} \pi_i f_i(\underset{\sim}{x};\ \underset{\sim}{\nu}). \tag{2.1}$$

Letting the vector $\underset{\sim}{\phi} = (\underset{\sim}{\pi}',\ \underset{\sim}{\nu}')'$ denote all the unknown parameters, the log likelihood of $\underset{\sim}{\phi}$ is given by

$$\log L(\underset{\sim}{\phi}) = \sum_{j=1}^{n} \log \left\{ \sum_{i=1}^{g} \pi_i f_i(\underset{\sim}{x}_j;\ \underset{\sim}{\nu}) \right\}. \tag{2.2}$$

Under the normality assumption,

$$\underset{\sim}{X}_j \sim N(\underset{\sim}{\mu}_i,\ \underset{\sim}{V}_i) \text{ in } \Pi_i \ (i=1,\cdots,g), \tag{2.3}$$

where the covariances are unequal, $L(\underset{\sim}{\phi})$ is unbounded, and so the maximum likelihood estimator (MLE) of $\underset{\sim}{\phi}$ does not exist (Kiefer and Wolfowitz, 1956). However, Kiefer (1978) verified that for p = 1 there is a sequence of roots of the likelihood equation,

$$\partial \log L(\underset{\sim}{\phi}) / \partial \underset{\sim}{\phi} = \underset{\sim}{0} , \tag{2.4}$$

which is consistent, asymptotically normal and efficient. With probability tending to one, these roots correspond to the local maximum of $L(\underset{\sim}{\phi})$. Unfortunately with mixture models, the likelihood equation (2.4) has multiple roots so there is the problem of which root to choose. This is discussed in some detail in Basford and McLachlan (1985d) and McLachlan and Basford (1987), and will be enlarged on later in this section.

With equal covariance matrices under model (2.3), so that

$$\underset{\sim}{X}_j \sim N(\underset{\sim}{\mu}_i, \underset{\sim}{V}) \text{ in } \Pi_i, \ (i=1,\cdots,g) \tag{2.5}$$

the MLE of $\underset{\sim}{\phi}$ does exist and is strongly consistent. It can be seen from Redner (1981) that the MLE is strongly consistent in the case where attention is restricted to a compact subset of the parameter space. But Basford and McLachlan (1985d) reported that Perlman noted that the strong consistency of the MLE, even for the unrestricted

(non-compact) parameter space, follows here from Kiefer and Wolfowitz (1956). Their conditions require that the mixture density converges, though not necessarily to zero, as $\underset{\sim}{\phi}$ tends to the boundary of the parameter space.

Under (2.3), the likelihood equation (2.4) is equivalent to

$$\hat{\pi}_i = \sum_{j=1}^{n} \hat{\theta}_{ij} \; / \; n, \quad (i=1,\cdots,g) \tag{2.6}$$

$$\hat{\underset{\sim}{\mu}}_i = \sum_{j=1}^{n} \hat{\theta}_{ij} \, \underset{\sim}{x}_j \; / \; n \, \hat{\pi}_i, \quad (i=1,\cdots,g) \tag{2.7}$$

and

$$\hat{\underset{\sim}{V}}_i = \sum_{j=1}^{n} \hat{\theta}_{ij} (\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)(\underset{\sim}{x}_j - \hat{\underset{\sim}{\mu}}_i)' \; / \; n \, \hat{\pi}_i, \quad (i=1,\cdots,g). \tag{2.8}$$

In these equations, the posterior probability that $\underset{\sim}{x}_j$, (really the element with observation $\underset{\sim}{x}_j$), belongs to $\Pi_i$ is given by

$$\theta_{ij} = \pi_i f_i(\underset{\sim}{x}_j; \, \underset{\sim}{\nu}) \; / \; \sum_{u=1}^{g} \pi_u f_u(\underset{\sim}{x}_j; \, \underset{\sim}{\nu})$$

$$= \frac{\pi_i |V_i|^{-p/2} \exp\left\{-\frac{1}{2} (\underset{\sim}{x}_j - \underset{\sim}{\mu}_i)' \, \underset{\sim}{V}_i^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu}_i)\right\}}{\sum_u \pi_u |V_u|^{-p/2} \exp\left\{-\frac{1}{2} (\underset{\sim}{x}_j - \underset{\sim}{\mu}_u)' \, \underset{\sim}{V}_u^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu}_u)\right\}} \tag{2.9}$$

and $\hat{\theta}_{ij}$ is the value of $\theta_{ij}$ with $\underset{\sim}{\phi}$ replaced by $\hat{\underset{\sim}{\phi}}$. Then $\underset{\sim}{x}_j$ is assigned to $\Pi_u$ if

$$\hat{\theta}_{uj} > \hat{\theta}_{ij} \ , \ (i=1,\cdots,g;\ i \neq u). \tag{2.10}$$

Equations (2.6) to (2.9) can be solved iteratively and these iterative estimates $\underset{\sim}{\phi}^{(q)}$ can be identified with those obtained by directly applying the EM algorithm of Dempster, Laird and Rubin (1977). Then provided $L(\underset{\sim}{\phi})$ is bounded above, $L(\underset{\sim}{\phi})$ converges to some local maximum $L^*$, provided the sequence is not trapped at some saddle point (Wu, 1983; Boyles, 1983). With mixture models, the likelihood often has multiple maxima, and so the EM algorithm should be repeated for several different starting values of $\underset{\sim}{\phi}$. If the data is univariate or bivariate, then a simple plot should enable suitable choices of starting allocations of the elements into g groups. This then provides an initial estimate of the unknown parameter vector $\underset{\sim}{\phi}$. When the data have more than two measurements on each element, probably the simplest procedure is to apply one of the hierarchical techniques and use its resulting allocation at the g group level as the starting allocation for the mixture method.

Basford and McLachlan (1985d) discuss the choice of suitable starting values during the search for all local maxima, as well as the problem of which of these to choose.

The obvious choice (Everitt, 1984) for the root of the likelihood equation is the one corresponding to the largest of the local maxima (assuming all have been located), although it does not necessarily follow that the consequent estimator is consistent and asymptotically efficient; see Lehman (1980, page 234). However, it would appear from the results of Hathaway (1983) that the estimator of $\phi$ corresponding to the largest of the local maxima is consistent and efficient at least for p=1. As noted earlier, the MLE does exist under the homoscedastic model (2.5) and is strongly consistent. Hence, from Lehman (1983, page 421), the choice of the largest of the local maxima is straightforward here assuming of course that the homoscedastic model is appropriate.

## 3. COMPARABLE METHODS FROM SAS (CLUSTER)

To provide a base for comparison, two of the SAS clustering methods with the most similar assumptions to the mixture maximum likelihood method were chosen. These hierarchical procedures are:

(i) Ward's method

(ii) EML method.

Both of these methods join clusters to maximize the likelihood at each level of the hierarchy assuming the data were sampled from a multivariate normal mixture of underlying populations with equal spherical covariance matrices. The difference between them is that Ward's minimum variance

method assumes equal sampling probabilities of the underlying populations while the EML method assumes unequal sampling probabilities. Hence Ward's method tends to join clusters with a small number of elements and is strongly biased toward producing clusters of roughly the same size whereas the EML method is somewhat biased toward unequal-sized clusters (SAS User's Guide: Stastistics; Version 5 Edition, 1985). Ward's method was put forward by Ward (1963) while the EML method was derived by W.S. Sarle of SAS Institute Inc., from the maximum likelihood formula obtained by Symons (1981, page 37, eq.[8]) for disjoint clusters.

Both methods are hierarchical agglomerative procedures in that they start with n clusters (or groups) each of one element and progressively fuse until a single cluster of n elements is obtained. As explained in Williams (1976), hierarchical procedures optimize a route between these two extremes. He noted that agglomerative strategies suffer from two disadvantages, the first of which is computational. The user's interest is normally concentrated in the higher levels of the hierarchy, so that it is almost invariably necessary to establish the complete hierarchy from individual elements to a single group of all elements. Secondly, an agglomerative system is inherently prone to a small amount of misclassification, the ultimate cause of which is that the process begins at the inter-individual level, where the possibility of this type of error is greatest. Once a cluster has been formed from a previous fusion in the hierarchy, it cannot be broken.

From the user's viewpoint, both methods have the advantage of providing a unique solution at any given level of the hierarchy. Also, they are not iterative techniques, so no starting values are required. This is because, at each level of the hierarchy, they work from the clusters obtained at the previous level and only two clusters are joined at any one time. However, it must be considered whether the stated assumptions are appropriate. Consider the requirement of equal spherical covariance matrices. This assumes, not only that (2.5) holds, but that

$$\underset{\sim}{V} = \sigma^2 \underset{\sim}{I} \qquad (3.1)$$

where $\underset{\sim}{I}$ is the identity matrix and $\sigma^2$ is not specified (Anderson, 1958, page 260). Thus the attributes measured on the elements are independent, the variances of these attributes are equal, and this covariance matrix is common to all of the underlying populations. This appears to be rather an unrealistic restriction unless all of the data have been prestandardized. Federer, McCulloch and Miles-McDermott (1986) have expressed grave reservations about such standardization when applying the multivariate technique of principal components. One of the aims of cluster analysis is to establish clusters which will enable a better perception and understanding of the information obtained on the elements, by observing the structure and relativities of the clusters. Hence it would not seem to be in the best interest

of the user to bias the results by any prestandarization.

## 4. DESCRIPTION OF THE DATA SETS

The first data set to be considered is the well known Iris data published by Fisher (1936). It consists of four measurements (sepal length, sepal width, petal length and petal width) on 50 plants from each of three species of Iris: *Iris setosa*, *Iris versicolor* and *Iris virginica*; denoted here by $\Pi_1$, $\Pi_2$ and $\Pi_3$, respectively. Many clustering techniques have been applied to this data set (see, for example, Kendall, 1966; Friedman and Rubin, 1967; Scott and Symons, 1971; Basford and McLachlan, 1985a). Hawkins' (1981) test of normality and homoscedasticity indicated normality with heteroscedasticity (see Fatti, Hawkins and Raath, 1982, page 64). It is worth noting , however, that some other analyses (Small, 1980; Royston, 1983) cast doubt on the normality of this data set.

The second data set under consideration was taken from Habbema, Hermans and van den Broek (1974), where, in the context of genetic counselling, the question of discriminating between normal women and haemophilia A carriers was considered on the basis of two variables ($\log_{10}$ (AHF activitity) and $\log_{10}$ (AHF-like antigen)). Reference data on 30 observations on known non-carriers (or normals) and 45 observations on known obligatory carriers were available with these populations denoted by $\Pi_1$ and $\Pi_2$ respectively. Note that the data set is to be analysed in a

general context, as it was in Basford and McLachlan (1985d), and not in its original setting where each observation $\underset{\sim}{x}_j$ would have an associated prior probability of being a non-carrier, taken to be the genetic chance of being normal as ascertained from the pedigree of the individual. Basford and McLachlan (1985d) showed that Hawkins' (1981) test indicated normality with heteroscedasticity. Detailed inspection showed that the only possible source of the indicated heteroscedasticity is in the difference in the sample variance of the second component ($\log_{10}$ (AHF-like antigen)). It will be seen that the adoption of a heteroscedastic model is of some consequence here as the clustering of the sample under this model is much more effective than that obtained under the assumption of homoscedasticity.

## 5. APPLICATION

For each of the two real data sets, the clustering methods were applied assuming that the data were a sample from a mixture of underlying multivariate normal populations. All results are presented at the group level correponding to the known number of underlying populations; this being g=3 for the Iris data and g=2 for the haemophilia A data. Only a single analysis is required for each of the hierarchical techniques but the mixture maximum likelihood method, applied for both models (2.5) and (2.3), needs to be run from several different starting allocations in an attempt to determine all

local maxima. As indicated in Section 2 the resulting groupings from the hierarchical techniques, as well as the known true grouping of the elements, were used as initial allocations for the Iris data. Basford and McLachlan (1985d) have a detailed discussion on initial allocation selection for the haemophilia A data. In addition to those mentioned there, the resulting groupings from the hierarchical techniques were also used. Such initial allocations of the elements into groups enable estimation of the unknown parameters in the model and hence provide a reasonable start for the iterative procedure of the EM algorithm. For the mixture maximum likelihood method, applied here using the program KMM (McLachlan and Basford, 1987), the grouping corresponding to the largest of the local maxima is presented as the solution.

The four clustering methods, in order of decreasing restrictions, were as follows:

(i) Wards' method - equal spherical covariance matrices and equal sampling proportions,

(ii) EML method - equal spherical covariance matrices and unequal sampling proportions,

(iii) KMM - equal covariance matrices and arbitrary sampling proportions,

(iv) KMM - arbitrary covariance matrices and arbitrary sampling proportions.

To explain the method of presentation of the results in Table 1, consider the first small table corresponding to the application of Wards' method to Fisher's Iris data. The columns correspond to the known population of origin of the elements; each column total must equal 50. The rows correspond to the resulting allocation of the elements into three groups; here the 50 *setosa* plants had 15 of the *virginica* plants grouped with them, while one of the *versicolor* plants was grouped with the remaining *virginica* plants. The off-diagonals indicate the number of elements incorrectly allocated into groups; 16 in this case. Thus the smaller this number, the more accurate the result of the clustering technique.

## 6. DISCUSSION

Ward's minimum variance method is probably the clustering method of choice for many occasional users; it was the default option in SAS (CLUSTER) for some time. At each level of the hierarchy, it combines clusters in such a way that there is a minimum increase in the within group variance. This has particular intuitive appeal although if homogeneity of groups was of prime importance, then the non-hierarchical strategies should take precedence as then it is the structure of the individual groups which is optimized (Williams, 1976).

There is no doubt that the hierarchical procedures are much easier to use but they have recently been criticized in

the statistical literature. Hawkins, Muller and ten Krooden (1982, page 253) commented that most writers on cluster analysis "lay more stress on algorithms and criteria in the belief that intuitively reasonable criteria should produce good results over a wide range of possible (and generally unstated) models". They strongly support the increasing emphasis on a model based approach. Similarly, Aitkin, Anderson and Hinde (1981) felt mixture models were an appropriate and useful tool as "when clustering samples from a population, no cluster method is *a priori* believable without a statistical model". Also, as they pointed out, "cluster methods based on such mixture models allow estimation and hypothesis testing within the framework of standard statistical theory".

In fact, Ward's method does have an underlying mixture model acting at each level of the hierarchy but it is quite restrictive because of the assumption of equal spherical covariance matrices and equal sampling proportions. The EML method does relax the latter assumption by specifying unequal sampling proportions but it still only maximizes the likelihood at each level of the hierarchy where no existing cluster can be split. Hence it would seem more reasonable to employ a non-hierarchical procedure in which there is much more flexibility in combining elements into clusters at any desired group level. This is supported by the results in Table 1 where it is clear that the normal mixture maximum likelihood method of clustering produces fewer

misclassifications than either of the hierarchical methods.

It should be pointed out that SAS (CLUSTER) does not have a test for sphericity of the estimated common covariance matrix as this would be an unrealistic and unnecessary addition to each level of the hierarchy. However, it must be remembered that this is an assumption of both Ward's method and the EML method and perhaps should be investigated at the particular level of the hierarchy of interest to the user. It should also be noted that there is no test for homoscedasticity within the KMM program (McLachlan and Basford, 1987) either. Hawkins' (1981) test for multivariate normality and homoscedasticity was applied to the known population of origin of the data using a separate computer program. Since, in practice, the true origin of each observation is unknown, the sample must be clustered first, using the mixture approach for normal densities with unequal covariance matrices. The test is then applied to these clusters as if they were the true groups with no misclassifications. This is rather a crude approach, but according to Fatti, Hawkins and Raath (1982), it appears to work fairly well.

As expected with the equal group size in Fisher's Iris data, Ward's method gave a more correct allocation than the EML method. It is somewhat surprising then, that for Habbema's haemophilia A data with unequal group sizes, both methods had the same number of misclassifications. Although the numbers in the relevant tables in Table 1 are the same,

there were actually eight elements from $\Pi_2$ which were allocated differently under these two hierarchical techniques. In practice, the EML method took fifty times as much computer time as Ward's method. The results in Table 1 do not suggest, at least for the two data sets under study, any real advantage of this method over Ward's method.

As stated in Section 5, many starting allocations were used for each of the data sets in order to identify all local maxima of the likelihood for the non-hierarchical mixture method of clustering. For the Iris data, only the one maxima was obtained under each model (2.5) and (2.3), as convergence was always to the same solution of the likelihood equation. However for the haemophilia A data, Basford and McLachlan (1985d) identified three local maxima under the homoscedastic model (2.5), and two local maxima under the heteroscedastic model (2.3). The grouping corresponding to the largest of the local maxima has been used to provide the results quoted in Table 1. Basford and McLachlan (1985d) showed that under the inappropriate homoscedastic model, the solution of the likelihood equation gives a poor estimate of the proportions in which $\Pi_1$ and $\Pi_2$ are represented in the sample. As a consequence, it gives a worse allocation than the solution corresponding to the second largest of the local maxima located. Then the clustering of the data, as displayed in Table 2, is almost the same as that under heteroscedasticity with just one additional member of $\Pi_2$ misallocated. Basford and McLachlan (1985d) discuss the sensitivity of the

iterative process in the mixture maximum likelihood method to starting values, in particular for the mixing proportion parameter. Unfortunately, there does not appear to be a simple answer to this problem. It is suggested that if heteroscedasticity is suspected, then it would be advisable to use an appropriate heteroscedastic model.

Finally it should be remembered that for the results considered in this paper, only the final allocation of the elements into groups has been considered. Under the mixture maximum likelihood method of clustering, the estimated posterior probability of group membership is also given. It is worthwhile investigating these estimates as they indicate the degree of certainty with which an element belongs to a particular group. For instance, if the estimated posterior probability of an element being assigned to a particular population, according to the allocation rule (2.10), was 0.6 or even less then the user is aware of some uncertainty in stating that the element belongs to this group. With the solution of the likelihood equation under the mixture model, there is no insistence on outright allocation of the elements to the groups at each stage of the interative process. Hence, providing regularity conditions hold, the estimates so obtained have the desirable large sample properties of likelihood estimators; for example, consistency, asymptotic efficiency and normality.

Table 1

Results of applying the clustering methods to
(i) Fisher's Iris data
(ii) Habbema's haemophilia A data.

| ıstering Method | Fisher's Iris data | | | | Habbema's haemophilia A data | | |
|---|---|---|---|---|---|---|---|
| | $\Pi_1$ | $\Pi_2$ | $\Pi_3$ | | $\Pi_1$ | $\Pi_2$ | |
| rd's method | 50 | | 15 | | 27 | 27 | |
| | | 49 | | | 3 | 18 | (30) |
| | | 1 | 35 | (16) | | | |
| | $\Pi_1$ | $\Pi_2$ | $\Pi_3$ | | $\Pi_1$ | $\Pi_2$ | |
| I. ..ethod | 50 | | | | 27 | 27 | |
| | | 27 | 1 | | 3 | 18 | (30) |
| | | 23 | 49 | (24) | | | |
| | $\Pi_1$ | $\Pi_2$ | $\Pi_3$ | | $\Pi_1$ | $\Pi_2$ | |
| M (equal | 50 | | | | 30 | 25 | |
| | | 48 | 1 | | | 20 | (25) |
| variance matrices) | | 2 | 49 | (3) | | | |
| | $\Pi_1$ | $\Pi_2$ | $\Pi_3$ | | $\Pi_1$ | $\Pi_2$ | |
| M (arbitrary | 50 | | | | 27 | 12 | |
| | | 45 | | | 3 | 33 | (15) |
| variance matrices) | | 5 | 50 | (5) | | | |

Table 2

Result corresponding to the second largest local maxima when applying KMM with equal covariance matrices to Habbema's haemophilia A data.

| | $\Pi_1$ | $\Pi_2$ |
|---|---|---|
| | 27 | 13 |
| | 3 | 32 |

(16)

## REFERENCES


Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society* A 144, 419-461.

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley and Sons.

Basford, K.E. (1985). *Cluster analysis via normal mixture models*. Unpublished Ph.D. thesis, University of Queensland.

Basford, K.E. (1986). An introduction to the use of mixture models in clustering. Cornell University Biometrics Unit Technical Report BU-920-M, Ithaca, New York.

Basford, K.E. and McLachlan, G.J. (1985a). Estimation of allocation rates in a cluster analysis context. *Journal of the American Statistical Association* 80, 286-293.

Basford, K.E. and McLachlan, G.J. (1985b). Cluster analysis in a randomized complete block design. *Communications in Statistics - Theory and Methods* 14, 451-463.

Basford, K.E. and McLachlan, G.J. (1985c). The mixture method of clustering applied to three-way data. *Journal of Classification* 2, 109-125.

Basford, K.E. and McLachlan, G.J. (1985d). Likelihood estimation with normal mixture models. *Applied Statistics* 34, 282-289.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39, 1-38.

Everitt, B.S. (1984). Maximum likelihood estimation of the parameters in a mixture of two univariate normal distributions; a comparison of different algorithms. *The Statistician* 33, 205-215.

Fatti, L.P., Hawkins, D.M. and Raath, E.L. (1982). Discriminant Analysis. In *Topics in Applied Multivariate Analysis*, Ed. D.M. Hawkins, Cambridge: Cambridge University Press, 1-71.

Federer, W.T., McCulloch, C.E. and Miles-McDermott, N.J. (1986). Illustrative Examples of Principle Component Analysis. Cornell University Biometrics Unit Technical Report BU-920-M, Ithaca, New York.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.

Friedman, H.P. and Rubin, J. (1967). On some invariate criteria for grouping data. *Journal of the American Statistical Association* 62, 1159-1178.

Habbema, J.D.F., Hermans, J. and van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. *Compstat 1974: Proceedings in Computational Statistics*. Vienna: Physica-Verlag, 101-110.

Hathaway, R.J. (1983). Constrained maximum likelihood estimation for normal mixtures. *Computer Science and Statistics: The Interface*. 263-267.

Hawkins, D.M. (1981). A new test for multivariate normality and homoscedasticity. *Technometrics* 23, 105-110.

Hawkins, D.M. , Muller, M.W. and ten Krooden, J.A. (1982). Cluster analysis. In *Topics in Applied Multivariate Analysis*. Ed. D.M. Hawkins, Cambridge: Cambridge University Press, 303-356.

Hernandez-Avila, A. (1979). *Problems in Cluster Analysis*. Unpublished D. Phil. Thesis, University of Oxford.

Kendall, M.G. (1966). Discrimination and Classification. In *Multivariate Analysis*. Ed. P.R. Krishnaiah, New York: Academic Press, 165-185.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* 27, 887-906.

Kiefer, N.M. (1978). Discrete parameter variation: Efficient estimation of a switching regression model. *Econometrika* 46, 427-434.

Lehman, E.L. (1980). Efficient likelihood estimators. *The American Statistician* 34, 233-235.

Lehman, E.L. (1983). *Theory of Point Estimation*. New York: John Wiley and Sons.

McLachlan, G.J. and Basford, K.E. (1987). *Mixture Models: Inference and Applications to Clustering*. To be published in New York by Marcel Dekker.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Annals of Statistics* 9, 225-228.

Royston, J.R. (1983). Some techniques for assessing multivariate normality based on the Shapero-Wilk W. *Applied Statistics* 32, 121-133.

Scott, A.J. and Symons, M.J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* 27, 387-397.

Small, N.J.H. (1980). Marginal skewness and kurtosis in testing multivariate normality. *Applied Statistics* 29, 85-87.

Symons, M.J. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics* 37, 35-43.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58, 236-244.

Williams, W.T. (1976). Types of classification. In *Pattern Analysis in Agricultural Science*. Ed. W.T. Williams, Amsterdam: Elsevier Scientific Publishing Company, 76-83.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* 11, 95-103.